4C1-4

ハブ・オーソリティモデルによる 主要スポット・代表ユーザー抽出法

Identifying Key Spots and Representative Users by Hub and Authority Models

鈴木 優伽 *1　　斉藤 和巳 *1
Yuka Suzuki　　Kazumi Saito

*1静岡県立大学経営情報学部
School of Management and Information, University of Shizuoka

We address a problem of identifying key sightseeing spots from movement of people. To this end, after organizing the move points to some areas by the mean-shift clustering method, we propose to construct a network whose nodes and directed links correspond to the areas and movement between them, respectively, and to identify the key spots as some nodes of the network indicated by some ranking methods such as PageRank. In our experiments using three datasets, we show that our method is vital and promising.

1. はじめに

近年，2020 年の東京オリンピックの開催決定や外国人観光客の増加を背景に観光産業に大きな期待が寄せられている．観光産業を活性化させるためには，旅行者がどのような観光地に行き易いのか，すなわち，旅行者の行動パターンや，主要な観光地の特徴などを明らかにした上で，新たな観光戦略を構築する必要がある．従来，旅行者の行動パターンの把握や，主要な観光地の抽出・分析などの調査は，アンケートによる紙面調査や，飛行機，列車などの旅行者の流入量，宿泊施設の稼働率といった統計調査によって行われていた．しかしながら，アンケートによる紙面調査は旅行者の負担が大きく，記入漏れ，時刻情報が不確実であるなど，正確性に欠ける可能性も否定できない．また，旅行の形態も団体から，個人・家族・友人などの少人数での旅行へと変化が見られ，よりミクロスケールな分析が重要であると考えられる．

本稿では，そのような背景を踏まえ，米国が構築した全地球測位システム (Global Positioning System: 以下，GPS) を利用し，複雑ネットワークの分析手法に基づいた，旅行者の行動パターンや主要な観光地の抽出・分析を行う．具体的には，従来研究 [Arase 10, Cao 10, Xin 10] と同様に，オンライン写真共有サイト Flickr(http://www．flickr．com) に投稿された写真に付随する GPS データを利用する．収集したデータを基に，インターネット上の Web ページの重要度指標である PageRank スコアを利用し，重要度指標の高い観光地の抽出や，旅行者の行動パターンなどを明らかにする．GPS ログデータから，複雑ネットワークの観点を基に，主要な観光地の抽出や旅行者の行動パターンの分析を行っている研究として文献 [Zheng 09] が挙げられる．そこでは，定められた閾値の中で作成された GPS ログデータ集合をクラスタリングした後，各クラスタをノードとして扱い，Hits スコアを用いることで分析を行っている．

一般的に，旅行者は一部のスポットだけに滞在するのではなく，一定のエリアのスポットに滞在すると考えられる．すなわち，あるスポットの重要度ではなく，あるエリアの重要度が参考にできれば，より多くの旅行者・観光エリアの有益性を高めることが出来る．そのため，mean-shift 法 [Crandall 09] を用いてクラスタリングした後，各クラスタをエリアとして扱い分析を行う．その際，PageRank スコアを用いることを本研究の提案手法とし，文献 [Zheng 09] のように HITS に基づく手法との結果の違いを検証する．また，各観光地を基にネットワークを構築した際の，構築したネットワークの性質や，スコア上位に抽出されるエリアの特徴，そこから考えられる旅行者の行動パターンについて分析する．

本稿の構成は以下の通りである．最初に提案手法について，2 章で詳細に説明する．次の 3 章では，実験設定について用いたデータと共に説明する．4 章で，実験結果への考察を述べ，最後に 5 章で本稿をまとめる．

連絡先: 鈴木優伽，静岡県立大学経営情報学部経営情報学科，静岡県静岡市駿河区谷田 52-1，054-264-5436，b12056@u-shizuoka-ken.ac.jp

2. 提案手法

本研究では，旅行者が Flickr に投稿した写真の位置情報を基に滞在エリアを決定する．ユーザーを $u \in U$，各ユーザーが撮影した写真の集合を $P_u = \{p_{u1}, p_{u2}, \cdots, p_{u|p_u|}\}$ とする．写真 p_{uk} は，緯度経度の位置情報，タイムスタンプの時刻情報を持ち，各々を $(p_{uk}.Lat)$, $(p_{uk}.Long)$, $(p_{uk}.T)$ とする．写真 p_{uk} のタイムスタンプを基に，ユーザー u が投稿した写真の位置集合を $L_u = \{l_{u,1}, l_{u,2}, \cdots, l_{u,|P_u|}\}$ とする．ただし，$l_{u,k} = (p_{uk}.Lat, p_{uk}.Long)$，$p_{u,k}.T \leq p_{u,k+1}.T$ である．また，ユーザー u の第 k 番目の写真の位置 l_{uk} を撮影地点とし，全撮影地点集合を $L = \bigcup_{u \in U} L_u$ とする．提案法は以下のステップで構成される．

1. 全撮影地点 L を mean-shift 法でクラスタリング
2. 各クラスタを滞在エリアとしネットワークを作成
3. 作成したネットワークで PageRank アルゴリズム実行

ステップ 1，2 について次節以降で詳細に説明する．

2.1 mean-shift クラスタリング

mean-shift は与えられたサンプル集合 X で定義されるカーネル密度関数における極大値探索法であり，それを利用したクラスタリング法が mean-shift クラスタリングである．本研究ではサンプル集合 $X = \{x_1, x_2, \cdots, x_N\} = L$ と定義し，カーネル関数としてガウシアンカーネル $\mathrm{K}(y, x_i) = \exp(-\|y - x_i\|/2\sigma^2)$ を用いた．ここで，y は任意の観測点を表わす．σ はカーネルサイズを決定するパラメータであり適宜設定する．また，本研究

では観測点 y における mean-shift ベクトル $m_{\sigma,\mathrm{K}}(y)$ を以下の式で定義する．

$$m_{\sigma,\mathrm{K}}(y) = \frac{\sum_{i=1}^{n} x_i \mathrm{K}(\|y - x_i)/\sigma^2)\|}{\sum_{i=1}^{n} \mathrm{K}(\|y - x_i)/\sigma^2)\|} - y \tag{1}$$

任意の観測点 y を出力とした際の，収束位置 y^c を求めるためのアルゴリズムは以下の通りである．また，mean-shift クラスタリングで得られるクラスタ集合を $V = \{v_1, v_2, \cdots\}$ とすると，c は V の要素の添え字に対応し y^cはクラスタ $v_c \in V$ の極大点となる．以下では，クラスタ $v^c \in V$ を滞在エリアとする．

Algorithm 1 mean-shift Procedure

Input: $X = \{x_1, x_2, \cdots, x_N\},\ x_i \in R^2$
where x_i is a two dimensional vector denoting Latitude and Longitude

1: Initialize $y_0 \leftarrow x_i,\ t = 1, y_1 \leftarrow m_{\sigma,K}(y_0)$
2: **while** $\|m_{\sigma,K}(y_t) - m_{\sigma,K}(y_{t-1})\| \geq$ threshold **do**
3: $\quad y_{t+1} \leftarrow m_{\sigma,K}(y_t)$
4: $\quad t \leftarrow t + 1$
5: **end while**
6: $y^c \leftarrow m_{\sigma,K}(y_t)$

Output: y^c

サンプル集合 X の各点 $x_i \in X$ に対し，以下のステップでクラスタリングを行う．

(St1) 各点 x_i に対し mean-shift Procedure を適応し，収束位置 x_i^c 計算;

(St2) 任意の 2 点 x_i, x_n の収束位置が閾値以下か判断;

(St3) $\|x_i^c - x_n^c\| \leq$ threshold ならば 2 点を同じクラスタに入れる ;

(St4) クラスタリングが終わるまで (St2)，(St3) を繰り返す;

2.2 ネットワーク作成

本節では，ネットワーク作成法について述べる．

ユーザー $u \in U$ の撮影地点 l_{uk}, l_{uk+1} が割り当てられた滞在エリアを $v_j = \mathrm{C}(l_{uk}), v_m = \mathrm{C}(l_{uk+1})$ とする．この時，滞在エリア v_j からノード集合 V，リンク $e(v_j, v_m)$ からリンク集合 E，多重度 $m(v_j, v_m)$ から多重度集合 M を構成し，多重有向ネットワーク $G = (V, E, M)$ を構築する．ここで，$e(v_j, v_m), m(v_j, v_m)$ は以下で定義される．ただし，本研究では自己リンクは考慮しないものとする．

$$e(v_j, v_m) = e(\mathrm{C}(l_{uk}), \mathrm{C}(l_{uk+1})); u \in U,\ \mathrm{C}(l_{uk}) \neq \mathrm{C}(l_{uk+1}) \tag{2}$$

$$m(v_j, v_m) = |\{(v_j, v_m) \mid u \in U, v_j = \mathrm{C}(l_{uk}), v_j = \mathrm{C}(l_{uk+1})\}| \tag{3}$$

例えば，ユーザー u がノード v_j に割り当てられた撮影地点 $\mathrm{C}(l_{uk}) = v_j$ で撮影した写真を投稿したのち，ノード v_m に割り当てられた撮影地点 $\mathrm{C}(l_{uk+1}) = v_m$ で撮影した写真を投稿したのならば，v_j, v_m 間にリンクを付与する．

3. 実験による評価

オンライン写真共有サイト Flickr に投稿された写真のうち，神奈川・京都・伊豆の 3 地域で投稿された写真データを用いた．写真データ数は神奈川・京都・伊豆で各，166,712・76,999・38,265 であり，ユーザー数はそれぞれ，5,206・3,980・1,597 である．クラスタリング後の各地域の滞在エリアは図 1 の通りである．また，本研究では比較手法として，ネットワークの中心性指標である入出次数と Hits スコアを用いる．

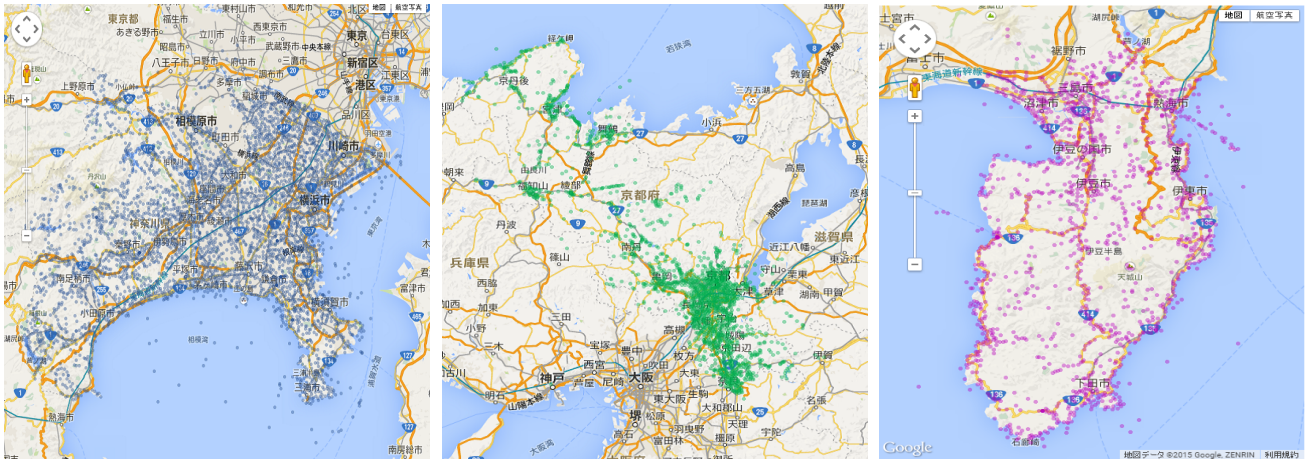
3.1 ネットワークの性質分析

本節では，各地域で作成されたネットワークの性質について分析する．図 2(a) は，神奈川・京都・伊豆の各地域の滞在エリアにおける撮影地点数の分布である．図中，青い点は神奈川，黄緑色の点は京都，ピンク色の点は伊豆を表す．図 2(a) を見ると，神奈川・京都・伊豆のどの地域においても，撮影地点数にスケールフリー性がみられる．図 2(b), 2(c) は，各地域のネットワークにおける近傍ノード数をプロットしたものである．ここで，in-neighbour は入リンクによる近傍ノード，out-neighbour は出リンクによる近傍ノードである．図 2(b), 2(c) を見ると図 2(a) と同様，近傍ノード数の分布にもスケールフリー性がみられる．すなわち，一部のエリアは多くの撮影地点を持つが，大多数のエリアは少数の撮影地点しか持たず，また，一部の旅行者は多くのエリアを訪れるが，大多数の旅行者は一部のエリアにしか訪れないという事が確認できる．

また，図 3(a) は各地域のネットワークにおける，入りリンクによる近接ノード数とその時の多重度をプロットしたものであり，図 3(b) は，出リンクによる近接ノード数とその時の多重度をプロットしたものである．図 3(a), 3(b) を見ると，多数のエリアと近接しているが，その繋がりが弱いエリアが存在することや，少数のエリアと近接しており，その繋がりが強いエリアが存在することを確認できる．これらのことから，クラスタをノードとして構成したネットワークも，通常の複雑ネットワークと同様の性質を持っていると考えられる．

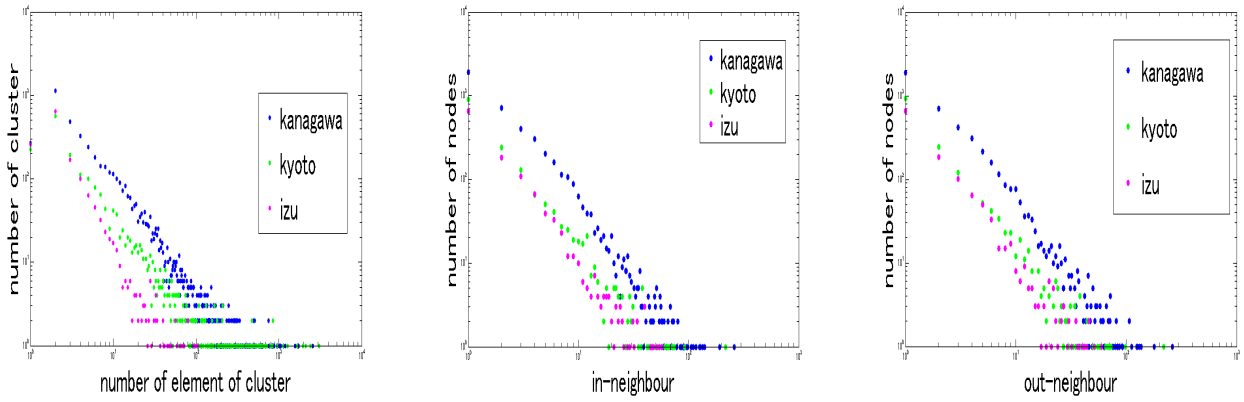
3.2 重要エリア

本節では，PageRank スコア, Hits スコア，入出次数が上位のエリアを重要エリアとし,，各指標で抽出されたエリアの特徴や違いについて分析する．ここでは紙面の都合上，伊豆地域での抽出エリアに焦点を絞り考察を述べていく．また，表に各指標でのスコア上位 5 エリアを示す．表 1 を見ると，駅や，ペリーロードといった道，熱海市銀座町といった駅周辺のエリアは Hits スコア・入出次数が高い事を確認できる．一般に，旅行者が移動手段として利用する駅や，道，駅周辺のエリアは，旅行におけるスタートエリア・中間地点エリア・ゴールエリアといったゲートウェイ的な存在であるために，相互リンクが多くなり，Hits スコアや入出次数が高くなると予想できる．そのため，今回の抽出結果は妥当な結果であると考えられる．また，PagaRank スコア上位エリアをみると，Hits スコア・入出次数では抽出されなかった沼津港・沼津深海水族館や，修善寺といった観光地エリアが抽出されている．これは，PageRank スコアの,「多くの重要なノードからリンクを張られているノードは重要である」という考えに沿った結果であると考えられる．すなわち，多くの旅行者が利用する駅や駅周辺のエリアから移動される・移動するエリアは，旅行者にとって旅行の目的地エリアとなる重要なエリアであると考えられ，観光地エリアが抽出されるのは自然な結果である．このことから，PageRank スコアを用いることで，Hits や入出次数といった指標で得られない，目的地的エリアが重要エリアとして抽出できることが確認できる．つまり，リンクの多重度や数を重要とする Hits・入出次数では，人の出入りが多い駅や駅周辺といったゲートウェイ的エリアが抽出されてしまうが，PageRank スコアを用いるという提案手法を用いることで，観光地などの目的地エリアを抽出でき，提案法の有効性が示唆される．



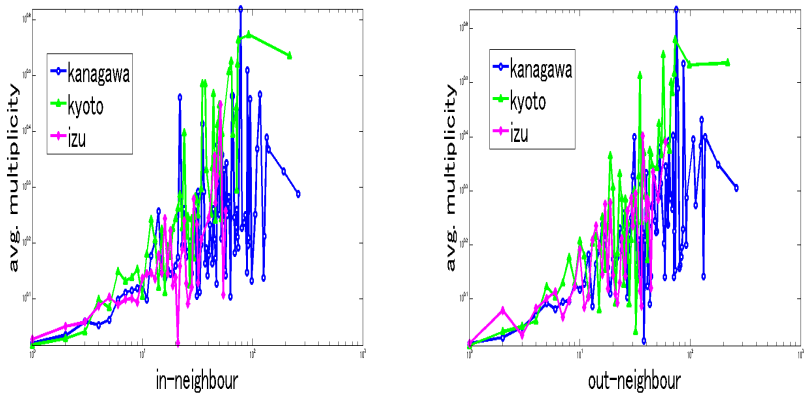
(a) 神奈川　(b) 京都　(c) 伊豆

図 1: 各地域の滞在エリア



(a) リンク次数分布　(b) 入ノード数分布　(c) 出ノード数分布

図 2: 各地域ネットワークの分布



(a) 入ノード数との多重度相関　(b) 出ノード数との多重度相関

図 3: 各地域ネットワークの多重度相関

表 1: 伊豆で抽出された重要スポット

Rank	in-Degree	out-Degree	Hub-score	Authority-score	in-PageRanke-score	out-PageRank-score
1	熱海駅	熱海駅	三島駅	熱海駅	熱海駅	沼津港
2	三島駅	ペリーロード	熱海駅	ペリーロード	三島駅	三島駅
3	熱海市銀座町 8	三島駅	下田駅	下田駅	沼津港	熱海駅
4	ペリーロード	下田駅	修善寺	三島駅	修善寺	ペリーロード
5	下田駅	熱海市銀座町 8	熱海市銀座町 8	白浜	下田駅	下田駅

表 2: 京都で抽出された重要スポット

Rank	in-Degree	out-Degree	Hub-score	Authority-score	in-PageRanke-score	out-PageRank-score
1	京都駅	京都駅	京都駅	京都駅	京都駅	京都駅
2	伏見稲荷大社	伏見稲荷大社周	伏見稲荷大社	伏見稲荷大社	伏見稲荷大社	伏見稲荷大社周
3	稲荷駅	稲荷駅	鴨川周辺	鴨川周辺	平等院鳳凰堂	平等院鳳凰堂
4	鴨川周辺	鴨川周辺	金閣寺	四条通り	稲荷駅	東寺周辺
5	四条通り	四条通り	四条通り	金閣寺	鴨川周辺	稲荷駅

表 3: 神奈川で抽出された重要スポット

Rank	in-Degree	out-Degree	Hub-score	Authority-score	in-PageRanke-score	out-PageRank-score
1	川崎駅	川崎駅	川崎駅	川崎駅	川崎駅	川崎駅
2	横浜駅	横浜駅	横浜駅	横浜駅	横浜駅	横浜駅
3	武蔵小杉駅	戸塚駅	横浜ダイヤビル	IKEA 港北	戸塚駅	横浜 LMT
4	戸塚駅	武蔵小杉駅	武蔵小杉駅	武蔵小杉駅	IKEA 港北	戸塚駅
5	横浜 LMT	逗子沿岸	横浜 LMT	横浜 LMT	武蔵小杉駅	センター北駅

3.3 多重度による旅行者の行動パターン分析

本節では，近接ノード数とその時の多重度に着目し，地域ごとの旅行者の行動パターンについて分析していく．図 3(a) を見ると，神奈川・京都・伊豆のどの地域においても，入りリンクによる近接ノード数が多くとも，多重度が高くなるとは限らないことが確認できる．特に，神奈川は，隣接ノード数の増加に伴う多重度の減少具合が顕著である．ここで，入りリンクによる隣接ノード数が多く多重度が高いというのは，多くの旅行者が滞在する人気をもち，周辺の滞在エリアと密に繋がっているエリアが存在することを示唆する．そのため，神奈川は多くの旅行者が滞在する人気エリアは存在するが，そのエリア間に密な繋がりが無いと考えられる．これは，神奈川が，旅行者が訪れやすい滞在スポットが横浜・鎌倉・箱根などのエリアに分散して存在しており，旅行者は横浜ならば横浜のエリア内のスポットしか滞在しないなどの行動パターンを持つためではないかと考えられる．また京都は，多重度の減少が神奈川ほど顕著に見られないが，これは，京都が，多くの旅行者が訪れやすい滞在スポットが京都市内のエリアに密集して存在しており，それらのエリア同士が近傍に存在しているため，旅行者が一定のエリアのみに長く滞在することなく，多くのエリアを滞在するという行動パターンをもつからではないかと考えられる．次に，図 3(b) をみると，京都・神奈川の出リンクによる隣接ノード数と多重度の関係は先ほど同様に，近接ノードの増加に伴い多重度が減少する傾向がみられたが，伊豆においては，近接ノードの増加に伴い多重度が増加する傾向がみられる．これは，伊豆に多くの旅行者が利用するゲートウェイ的エリアが明確に存在し，旅行者がそこから一定のエリアに移動する傾向を強くもつからだと考えられる．

4. おわりに

本研究では，一定の範囲のエリアをノードとした際の PageRank スコアを計算し，重要スポット抽出を行った．その結果，比較手法として用いた Hits や入出次数でとは異なるスポットの抽出が確認でき，提案法の有効性が示唆された．今後は，旅行者が撮影した時の天気，撮影者の性別などの属性値情報を考慮することで，より旅行者の需要にそったエリアを抽出することを目指す．

謝辞 本研究は, 総務省 SCOPE(No.142306004), 科学研究費補助金基盤研究 (C)(No.26330345) の補助を受けた．

参考文献

[Arase 10] Arase Y., Xie X., Hara T., and Nishio S.: ”Mining People’s Trips from Large Scale Geo-tagged Photos”, ACM Multimedia2010, pp. 133-142 (2010)

[Cao 10] Cao L., Luo J., Gallagher A., Jin X., Han J., Huang T.S.: ”A Worldwide Tuorism Recommendation System Based on Geotagged Web Photos”，ICASSP，pp.2274-2277 (2010）

[Crandall 09] Crandall D., Backstrom L., Huttenlocher D., and Kleinberg J.: ”Mapping the World’s Photos”, WWW2009, pp.761-770 (2009).

[Xin 10] Xin L., Changhu W., Jiang-Ming Y., Yanwei P., and Lei Z.: ”Photo2Trip: Generating Travel Route from Geo-Tagged Photos for Trip Planning”，ACM Multimedia2010, pp.143-152 （2010）

[Zheng 09] Zheng Y., Zhang L., Xie X., and Ma W.-Y., ”Mining Interesting Locations and Travel Sequences from GPS Trajectories”, WWW2009, pp. 791-800 (2009).